Evaluation Platform for Content-based Image Retrieval Systems

Petra Budikova, Michal Batko, and Pavel Zezula

Masaryk University, Brno, Czech Republic

Abstract. In all subfields of information retrieval, test datasets and ground truth data are important tools for testing and comparison of new search methods. This is also reflected by the image retrieval community where several benchmarking activities have been created in past years. However, the number of available test collections is still rather small and the existing ones are often limited in size or accessible only to the participants of benchmarking competitions. In this work, we present a new freely-available large-scale dataset for evaluation of content-based image retrieval systems. The dataset consists of 20 million high-quality images with five visual descriptors and rich and systematic textual annotations, a set of 100 test query objects and a semi-automatically collected ground truth data verified by users. Furthermore, we provide services that enable exploitation and collaborative expansion of the ground truth.

Keywords: large-scale image dataset, visual and textual annotation, ground truth, collaboration service

1 Introduction

Image search is a very attractive topic nowadays, as witnessed by the number of recent research papers and the rapid development of commercial image search systems. Still, a satisfactory solution seems to be yet a long way ahead due to a number of obstacles – the size of data, the semantic gap problem and a range of specific application issues. Many researchers are trying to overcome these problems using different ideas and approaches. To be able to pick out and develop the best ones, we need a common platform for testing and evaluation.

The lack of benchmarks for image search is a well-known problem. Although several freely-available test datasets exist, they do not cover all basic application areas. In particular, there is no large-scale dataset that could be used for evaluation of web image search, where scalability is one of the key issues. Therefore, we created a new extensive dataset and offer it hereby to other researchs.

Naturally, we need more than just the data to be able to compare the performance of various search methods. The other necessary part is the ground truth, i.e. the set of query objects and the respective sets of relevant result objects. Only then we can compare the precision, recall, and other metrics of the methods under examination. Unfortunately, creating the ground truth is a difficult

task for any dataset, let alone the very large ones. As there is no objective automatic way of deciding the relevance of image with respect to a given query image, we need to ask people to do the tedious task. In case of large datasets, it is hardly possible to manually examine all the images in the dataset. Therefore, we adopt a different approach and describe a way of creating a partial ground truth by pre-selecting a candidate set of a reasonable size that is then examined and refined by humans.

In addition to the dataset itself and the ground truth for a hundred different query images we also offer two web-service tools. The first tool allows evaluating any search method against the ground truth. The second tool provides an interactive user interface that allows to collaboratively create a new ground truth for additional images.

2 Related Work

In the early days of image retrieval, the Corel dataset was the first collection to be used for evaluation. It provided over 68,000 images, organized into classes of about 100 images, each with roughly the same topic. However, this artificial and relatively small dataset is not satisfactory as a benchmark nowadays. We need to take into account different applications, the data they use (scope, size, metadata available, etc.) and the user information-retrieval requirements. Serious efforts for building a complex benchmarking platform appeared in [15]. The proposed methodology was to be realized by the Benchathlon[1] project, where research groups were meant to cooperate on creating the testing platform. Unfortunately, this project does not seem to be making any progress. Another analysis of image evaluation campaign can be found in [8]. It describes the background of ImageEVAL competition, which took place in 2006. However, the only repeated and successful benchmarking activity we know of is the ImageCLEF[2] competition which has been running since 2003. Each year, the organizers define various challenges, provide data and topics and evaluate the submitted results.

Nonetheless, even the ImageCLEF activities are limited by the availability of *benchmark inputs*, as defined in [12]: the data collection (documents), the queries (topics) and the ground truth (relevance judgements). We review these three issues in more detail in the following sections. We mainly focus on large, general-purpose datasets, leaving aside specialized collections such as medical images, arts collections, etc.

2.1 Image databases

Gathering a large collection of image data is not a simple task due to the ownership and copyright issues. However, this can be overcome by using freely available web resources, such as the Flickr web gallery or Wikipedia. The following three datasets have been obtained this way. The first two have been composed to serve

[1] http://www.benchathlon.net/
[2] http://www.imageclef.org/

as the benchmarking sets and are used in the ImageCLEF competition. All of them provide both images and text metadata, but they differ in size, origins and scope of the metadata.

MIRFLICKR collection The MIRFLICKR collection[3] [9] consists of 1 million images (at the moment) downloaded from the social photography site Flickr. All images are available under a Creative Commons Attribution Licence. The images have been selected based on their high *interestingness* rating that is determined by factors such as where the click-throughs on the images are coming from, who comments on them, and whether they are marked as favorites. In addition, user-supplied Flickr tags, EXIF metadata and systematic image annotations are available. The visual descriptors provided are the MPEG-7 Edge Histogram and Homogeneous Texture descriptors, and the ISIS Group color descriptors.

Wikipedia collection The ImageCLEF 2010 Wikipedia collection[4] [18] extends the INEX MMWikipedia collection [19], which was created for the purpose of INEX evaluation campaign in 2007. Currently the collection consists of 237,434 Wikipedia images, their user-provided annotations, the Wikipedia articles that contain these images, and low-level visual features of these images. The collection was built to cover similar topics in English, German and French and it is based on the September 2009 Wikipedia dumps. Images are annotated in none, one or several languages. Image visual features include both local (bags of visual words) and global features (texture, color and edges). The collection is available for the participants of the ImageCLEF competition.

CoPhIR image set The CoPhIR dataset[5] [1] with 106 million processed images is currently the largest collection available for scientific purposes. It consists of metadata extracted from the Flickr photo sharing system. For each image, the collection contains a thumbnail image, a link to a corresponding entry at the Flickr web site, user-specified metadata (title, GPS location, tags, comments, etc.) and five MPEG-7 visual descriptors (Scalable Color, Color Structure, Color Layout, Edge Histogram and Homogeneous Texture). Since the data are not supervised, some of them are of poor quality – blurred or too dark/light images, images with sparse and erroneous annotations, different languages used in annotations, etc. While this may cause worse performance of search methods, the collection provides a good model of a real-world data.

2.2 Topics

The common goal of all search systems is fulfilling user's information need. Therefore, the test search topics should simulate what a real user of the system would instantiate as usage scenario. Furthermore, the volume (number) and diversity (variability) of the topics should cover the whole search domain and demonstrate statistical robustness of the results [12].

[3] http://press.liacs.nl/mirflickr/
[4] http://www.imageclef.org/2010/wiki
[5] http://cophir.isti.cnr.it

The usual ways of creating test search topics comprise a choice made by domain experts and an analysis of search system usage logs. In [18], the creation of topics for ImageCLEF 2010 Wikipedia Retrieval task is described in more detail. A candidate set of queries is derived from a search log file and topics from previous runs of the competition. From these, only such queries are accepted that have a sufficient number of relevant results in an organizers' search run.

Another issue is the query definition. Basically, there are three ways to go – query by example (image), query by text and query by both text and images. While query images are used in annotation tasks, they are not suitable for image retrieval since one image may often represent several concepts, while the imaginary user is only interested in one of them. Therefore, either complex text queries or images complemented by text are used in web-like image search tasks.

2.3 Ground truth

The ground truth data is used to decide the relevance of a result provided by a search system. In an ideal case, the ground truth should contain an indicator of relevance for each object in the dataset and each search topic. The relevance can be either binary (relevant, irrelevant) or expressed as a level of relevance, e.g. as a percentage. The relevance is decided by human judges, preferably more than one for each object and topic to balance the subjectivity of opinions.

Clearly, creating such a ground truth is a laborious effort. When only a few people are involved, be them domain experts or lab members, providing exhaustive relevance judgements is only feasible for relatively small datasets. For the large ones, some approximations are usually employed. The one that is mostly used in the evaluation campaigns is called *pooling*: only those objects are judged that appear among the top n images of any of the results submitted by the competitors [18]. However, this results in a one-time ground truth that cannot be meaningfully reused for evaluation of different result sets.

Alternatively, expert annotations can be provided for each image, using a defined categorization. This is also a tedious work, but only needs to be done once for each object in the dataset. Afterwards, a ground truth for a given query can be obtained by judging only objects that have the relevant keywords in their annotation. This approach has been adapted by the supervisors of the MIRFLICKR dataset [9].

The only way to obtain an exhaustive ground truth for large datasets is by employing many people. However, it is not easy to find the necessary motivation. One possible approach is to invest a considerable amount of money and pay for each judgement. This approach was adapted to create the ImageNet database [6], where the Amazon Mechanical Turk platform was used to manually clean a large set of candidate images. Another method is shown in the TagCaptcha image annotation system [13], where the authors propose to obtain annotations via the widely used Captcha challenge-response tests.

Altogether, it is obvious that it is difficult to create the ground truth data. The evaluation campaigns such as ImageCLEF gather relevance judgements during the competitions but these are not public in order to prevent cheating in

future competition runs. In consequence, there is a deficiency of ground truth data for testing outside the evaluation campaigns, which is definitely an obstacle in the development of new search methods.

3 Our dataset

Our objective is to provide a dataset that will enable to test systems for large-scale searching in terms of results quality (precision), efficiency (search time) and scalability. The important aspects are therefore (1) the size of the dataset, (2) its scope, and (3) the type of data provided. As to the size, the datasets that are used for benchmarking nowadays range in volume from hundred thousand to millions of images. We believe that even larger datasets are necessary to test the efficiency of methods for web searching. Regarding the scope, we are interested in a real-world dataset since the performance of search mechanisms is influenced by the distribution of objects in the domain. Finally, recent research [5, 10] indicates that the future of image searching seems to be in combining multiple modalities, typically visual features and text metadata describing the semantics. Therefore our dataset should contain at least these two modalities.

The Profimedia collection which we are offering to the research community satisfies all the discussed requirements. We obtained the image set from Profimedia[6], a web-site selling stock images produced by photographers from all over the world. The collection contains 20M high-quality images with rich and systematic annotations. For each image, we have extracted five MPEG7 [14] global visual descriptors recommended in [1]. Thus, each entry in the dataset consists of the following information:

- a thumbnail image;
- a link to the corresponding page on the Profimedia web-site;
- two types of image annotation: a title (typically 3 to 10 words) and keywords (about 20 keywords per image in average) mostly in English (about 95 %);
- five MPEG-7 visual descriptors extracted from the original image content: Scalable Color, Color Structure, Color Layout, Edge Histogram and Region Shape.

The dataset can be downloaded from **http://mufin.fi.muni.cz/profiset/** after registration and agreement to the usage terms. The data can be freely used for research purposes.

4 Query topics

When selecting the topics, we had the following requirements in mind: the queries should reflect real users' needs, the topics should be diverse both in content and in complexity, and there should be enough relevant results for each test query in the dataset.

[6] http://www.profimedia.com/

To achieve this we first created a set of candidate topics which comprised (1) popular queries from the search logs provided by Profimedia, and (2) several examples of queries that we know from experience to be either easy or difficult to process in content-based searching. Next, we run a top-30 query for each of the candidates using an aggregation of text and visual search (described in more detail in Section 5). Only the topics that had at least 10 relevant results were accepted into the final query set.

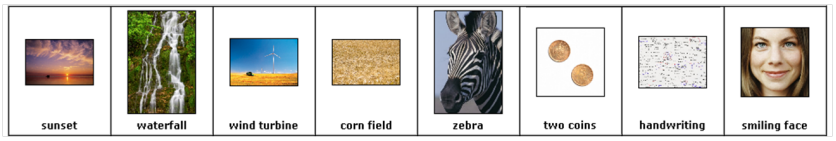

Fig. 1. Query object examples.

The test set contains 100 topics, each of which is defined by a single query image and a few keywords (typically one or two). The following categories are represented by the topics: activity (5 queries), animal (8), art (6), body part (5), building (3), event (3), food (8), man-made objects (16), nature (16), people (12), place (9), plant (2), specific building (4), and vehicle (3). Several examples of the query objects are shown in Figure 1.

5 Partial ground truth

As stated earlier, a full ground truth should contain relevance evaluation for each topic-object pair but creating it for a large dataset is only feasible when a lot of people are employed. Unfortunately, we lack the resources required to do such a job, so we have used the pooling approach and our lab colleagues acted as the judges. We are aware of the fact that the pooling approach can miss relevant images, thus the provided data represent a *partial ground truth*. However, we tried to create it in such a way that it should cover the majority of relevant images and we also provide tools for expanding the ground truth when needed.

In the benchmarking competitions, the pool of candidate images for the evaluation is usually composed of the top n submitted results. We applied a similar technique but we have used a set of our own search methods implemented over the MESSIF framework [3] that provided the results. These methods were designed in such a way as to retrieve as many different relevant objects as possible. Our query topics consist of image and text, thus we can work with these two modalities and apply text-based retrieval, content-based retrieval or a combination of both. Furthermore, many preprocessing (query expansion) and postprocessing (ranking) methods have been proposed recently to improve the search efficiency. As illustrated in Figure 2, we treat the query evaluation as a three-phase process, where each of the phases can be realized in several ways. The search methods we used to create the pool of candidate images are then formed by various combinations of the individual techniques.

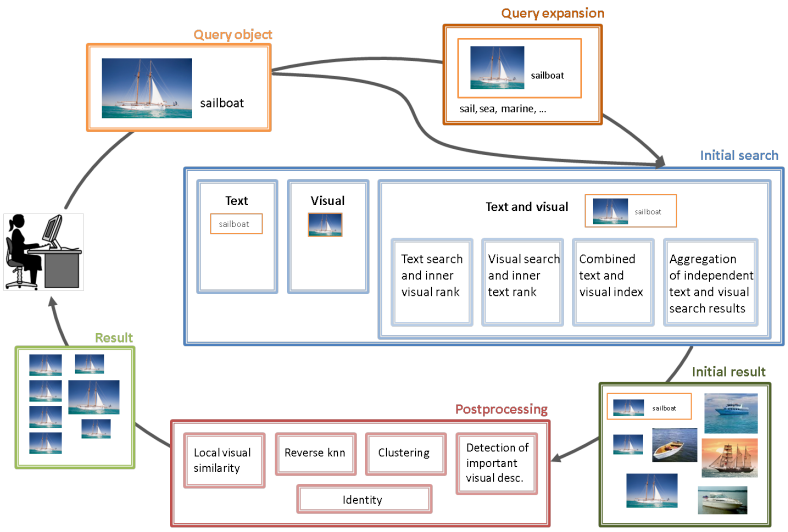

Fig. 2. The global search schema

5.1 Query expansion

The query expansion techniques [16] endeavour to automatically provide additional information to the query that will help to obtain better search results. Query expansion can be used to describe the user's information need more precisely (e.g. word sense disambiguation) or to overcome the gap between the query specification and the data available (e.g. automatic synonym expansion). For our test search methods, we chose the basic expansion technique that is often applied on short text queries. Using the WordNet [7] lexical database, we enriched the query with synonyms and hypernyms of the query keywords. This way, the relevant objects can be added into the candidate set even though their annotations are formulated differently.

5.2 Initial search

In the initial search phase, the query (expanded or not) is submitted to a search method which processes it over the whole dataset and produces an initial result set. We adopt three types of searching: text-based retrieval, content-based (visual) retrieval and combined text-and-visual search. All the functions are implemented using the MESSIF framework [3] and the MUFIN [17] system.

Text-based The text search is executed as a classical *tf-idf* retrieval, only with different weights used for keywords from user, keywords in the query image title and keywords in the image annotation.

Content-based The content-based search is based on the five MPEG-7 descriptors available and the respective distance functions as defined in the MPEG-7 standard [14]. The individual distances are combined using a weighted sum.

Text-and-visual The combined search aggregates text and visual similarity. We use a weighted sum aggregation function with three different settings of the respective weights. The combined search can be implemented in several ways:

- Text search and inner visual rank: The text search is run on a full database but all objects relevant by the text criterion are ranked by combined text-and-visual similarity and only then the top ranking results are returned.
- Visual search and inner text rank: Same as previous, only vice-versa.
- Combined text and visual index: The search is evaluated over a metric index structure that combines the text similarity and content-based similarity.
- Independent text and visual search: The two search methods are run separately and the ranked lists are aggregated. In [2] we described how the aggregation can be done efficiently in a distributed environment.

5.3 Postprocessing

The philosophy of postprocessing is based on the fact that the search engine can provide a result set one or two orders of magnitude larger than required with nearly the same costs. Additional evaluations of similarity can be computed over this initial result that would be too expensive to process over the whole dataset [4]. In our experiments, we applied the following ranking functions:

- Identity: No ranking is applied, the top objects from the initial result are displayed to the user.
- Identification of important visual descriptors: The variance of visual descriptors is analyzed over the initial result set, the descriptors with low variance receive higher weights.
- Clustering: Objects that are more similar to the other objects in the initial result are ranked higher than the outliers.
- Reverse kNN: The rank of an object is given as the number of objects that are more similar to it than the query object.
- Local visual similarity: local similarity is evaluated using the SIFT features [11], the top ranking objects are shown to the user.

5.4 Relevance judgement

Altogether with variable weights settings we created 140 search methods. For each query image, top-20 queries were evaluated by all methods and their merged results were displayed in a web interface shown in Figure 3. The judges were asked to mark each object as *very good*, *acceptable*, or *irrelevant*, which we transformed into relevance levels of 100 %, 50 % and 0 %, respectively. Using the numerical values, we evaluated the final relevance as an average of collected judgements.

5.5 Statistical evaluation

The ground truth data we obtained from our judges contain a considerable number of relevance evaluations which are a valuable resource for analysis of human

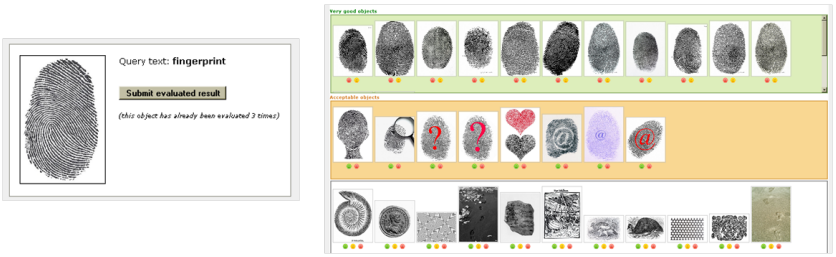

Fig. 3. The web interface for relevance evaluation

perception of similarity. In this section, we present several observations concerning both the properties of our dataset and the human factor in the evaluation.

The evaluation was performed by 15 participants, most of them students, graduates, or researchers in IT. Out of the 100 queries, each got evaluated at least twice, the total number of evaluations being 222. With the average number of candidate objects per query topic being 578, we obtained a total of 128,141 evaluated topic-object-user triplets. The evaluation process took a month, the actual time invested in the judgements being about 100 hours.

As mentioned earlier, we compute the relevance of a result object as the average of all evaluations we have for it. We find it suitable to categorize objects into the following categories: *perfect* (average relevance 100 %), *good* (at least 50 %), *partially relevant* (more than 0 %), and *irrelevant*. For each query topic in our testbed, there were in average 105 perfect result objects, 223 good objects and 315 irrelevant ones. However, the number of objects in each category differed considerably between individual queries – the lowest number of perfect results was 5 and 11 objects had less than 20 perfect results. The lowest number of good results per query was 53. We can conclude that our set of topics is suitable for testing as there are enough relevant objects to be found and, at the same time, enough queries with various difficulty levels are present (difficulty being inversely proportional to the number of relevant objects contained in the dataset).

When evaluating the results, the judges were not given any instructions on what shall be considered relevant. Therefore, their classification of results reflects their individual understanding of similarity and their expectations of image search system performance. While this is known to be subjective and inconsistent in different situations, all image retrieval systems are based on a tacit assumption that there exists some basic agreement in the individual opinions. Using our relevance evaluations, we can verify this assumption. The following table shows the percentage of identical evaluations, where all judges agreed on the (ir)relevance of a query object pair.

Number of evaluations	Identical evaluations	Unmatched (2 different)	Unmatched (3+ different)
2	80 %	20 %	–
3	70 %	27 %	3 %
4	73 %	21 %	6 %
5	65 %	20 %	15 %

For the sake of our ground truth it is also important to know whether two judgements (which we have for most queries) are sufficient to obtain a trustworthy relevance evaluation or whether more opinions are needed. Figure 4 shows how the percentage of objects with given relevance changes with the growing number of evaluations (we used the results with the most evaluations to obtain these graphs). We can observe that the results are quite stable, therefore the two judgements can be considered sufficient.

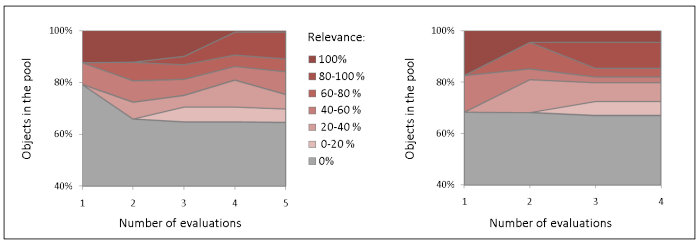

Fig. 4. The development of result evaluation

Finally, let us have a look on the methods we used to create the candidate pool. We employed a high number of combinations of search methods and post-processing techniques in order to discover as many relevant objects as possible. This approach has proved to be well suited as every combination did bring some relevant object to the results that was not found by any other method.

6 Provided functionality and Extensibility

In order to offer the tools created during the preparation of the partial ground truth for the research community's benefit, we have designed two web-services. The first one simplifies the benchmarking of an external search method against the existing ground truth. The other one allows to add a new image to the query set and collaboratively evaluate its partial ground truth.

Due to space limitations we only explain the services in general here. More details, the specifications, and the access to the services can be found on the dataset page **http://mufin.fi.muni.cz/profiset/** in the Services section.

6.1 Evaluation of external search method

Researchers proposing new search methods for image retrieval systems are welcome to use the Profimedia dataset as a benchmark. By downloading the dataset, the query set, and the ground truth, they can compute their own statistics on the effectiveness of their method. However, since our ground truth is only partial – given that it was evaluated from a limited set of candidate objects (see Section 5) – the new proposed method can be penalized on images that are relevant to the query object but were not included in the candidate objects.

To overcome this problem, the results of the new method (just the identifiers of the images) can be uploaded to our service. The service then checks all the objects that were in the original *candidate* set from which the ground truth was computed and any new image is presented via the web-interface. The user is then able to judge whether each of the new objects is *very good*, *acceptable*, or *irrelevant* in the same way as when the previous partial ground truth was created. Afterwards, the statistics of the new method using the updated partial ground truth are displayed.

Any such addition to the existing partial ground truth is also stored in our database and immediately available for download. Thus, the partial ground truth is collaboratively extended whenever a new method is tested via our service.

6.2 Additional query images

Since our query set consists of a hundred images while the dataset contains 20 million images, we offer a service that allows to evaluate the partial ground truth for an additional query image. In order to do that, we need a candidate set of images and then user judgements of the relevance of the respective images (as explained in Section 5).

Our service thus allows to upload a new query image (or select an existing image from the Profimedia dataset using its identifier). Then one or more candidate sets can be uploaded, e.g. retrieved by some new search methods (as in the other service above). Finally, the system asks whether the candidate set should be expanded by our search methods. We provide options for selecting our text-based, content-based, or combined methods as explained in Section 5. Since the candidate set creation is a computationally intensive task, a job is scheduled in our university GRID[7]. It can take some time until the candidate set is ready for the user judgement, so the service notifies the user by email.

Then, the new query object is available for the user evaluation via a web interface as shown in Figure 3. When at least one evaluation is complete, the query is available in the query set with the new partial ground truth. The query is then also offered for additional evaluations to other users.

7 Conclusions

In this paper, we present a new freely-available large-scale dataset for evaluation of content-based image retrieval systems. The dataset consists of 20 million high-quality images with five visual descriptors and rich and systematic textual annotations. For this dataset, we have prepared a set of 100 test query images from various categories and collected a partial ground truth for each of them. The partial ground truth was human-judged from candidate sets generated by 140 search methods. To allow exploitation and collaborative expansion of the ground truth, we offer two public web-services. The data and services are accessible on the web page **http://mufin.fi.muni.cz/profiset/**.

[7] http://www.metacentrum.cz/

Acknowledgments

This work has been partially supported by Brno PhD Talent Financial Aid and by the national research projects GAP 103/10/0886 and VF 20102014004. The hardware infrastructure was provided by the METACentrum under the programme LM 2010005.

References

1. Batko, M., Falchi, F., Lucchese, C., Novak, D., Perego, R., Rabitti, F., Sedmidubský, J., Zezula, P.: Building a web-scale image similarity search system. Multimedia Tools Appl. 47(3), 599–629 (2010)
2. Batko, M., Kohoutkova, P., Zezula, P.: Combining metric features in large collections. In: ICDE Workshops. pp. 370–377. IEEE Computer Society (2008)
3. Batko, M., Novak, D., Zezula, P.: MESSIF: Metric similarity search implementation framework. In: 1st DELOS Conference. LNCS, vol. 4877, pp. 1–10. Springer (2007)
4. Budikova, P., Batko, M., Zezula, P.: Similarity query postprocessing by ranking. In: 8th International Workshop on Adaptive Multimedia Retrieval (2010)
5. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40(2) (2008)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
7. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. The MIT Press (1998)
8. Fluhr, C., Moëllic, P.A., Hède, P.: Usage-oriented multimedia information retrieval technological evaluation. In: Multimedia Information Retrieval. pp. 301–306 (2006)
9. Huiskes, M.J., Lew, M.S.: The MIR Flickr retrieval evaluation. In: Proc. of the Multimedia Information Retrieval. ACM (2008)
10. Jain, R., Sinha, P.: Content without context is meaningless. In: ACM Multimedia. pp. 1259–1268 (2010)
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
12. Marchand-Maillet, S., Worring, M.: Benchmarking image and video retrieval: an overview. In: Multimedia Information Retrieval. pp. 297–300 (2006)
13. Morrison, D., Marchand-Maillet, S., Bruno, E.: TagCaptcha: annotating images with CAPTCHAs. In: Proc. of the ACM Multimedia. pp. 1557–1558 (2010)
14. MPEG-7: Multimedia content description interfaces. Part 3: Visual. ISO/IEC 15938-3:2002 (2002)
15. Müller, H., Müller, W., Marchand-Maillet, S., Pun, T., Squire, D.M.: A framework for benchmarking in CBIR. Multimedia Tools Appl. 21(1), 55–73 (2003)
16. Natsev, A., Haubold, A., Tesic, J., Xie, L., Yan, R.: Semantic concept-based query expansion and re-ranking for multimedia retrieval. In: ACM Multimedia. pp. 991–1000 (2007)
17. Novak, D., Batko, M., Zezula, P.: Generic similarity search engine demonstrated by an image retrieval application. In: Proceedings of SIGIR '09. p. 840 (2009)
18. Popescu, A., Tsikrika, T., Kludas, J.: Overview of the Wikipedia Retrieval Task at ImageCLEF 2010. In: CLEF (Notebook Papers/LABs/Workshops) (2010)
19. Westerveld, T., van Zwol, R.: The INEX 2006 Multimedia Track. In: INEX. pp. 331–344 (2006)